On Bayesian Learning from Bernoulli Observations

Pier Giovanni Bissiri[a,*], Stephen G. Walker[b]

[a]*Dipartimento di Statistica, Università degli Studi di Milano-Bicocca, Edificio U7, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy*
[b]*IMSAS, University of Kent, Canterbury, Kent, CT2 7NZ, UK tel: +44(0)1227 823800*

Abstract

We provide a reason for Bayesian updating, in the Bernoulli case, even when it is assumed that observations are independent and identically distributed with a fixed but unknown parameter θ_0. The motivation relies on the use of loss functions and asymptotics. Such a justification is important due to the recent interest and focus on Bayesian consistency which indeed assumes that the observations are independent and identically distributed rather than being conditionally independent with joint distribution depending on the choice of prior.

Keywords: Asymptotics, Kullback–Leibler divergence, Loss function.

1. Introduction

The aim of this paper is to provide a straightforward and concise justification of the Bayesian approach to updating probability beliefs, in the case of a Bernoulli sequence of random variables. It is then seen that the details of the result can be applied to general parametric families. The key is the use of a loss function combined with the notion of asymptotics. That is, a loss function on the space of probability distributions on $(0, 1)$ is employed which uses as information the prior knowledge and the observations. The general setting is made precise by appealing to obvious asymptotic requirements for the solution to the minimization of the loss function. Indeed, interest in Bayesian consistency has grown in the last years. See, for instance, Xing and Ranneby (2009) and references cited in this paper.

The use of loss functions is limitless within the world of applied sciences, no more so than within the decision sciences which includes statistics and particularly Bayesian statistics (Berger, 1993). To set the scene, if $\mathcal{A}$ is a set of actions, and the loss incurred is $L(a, X)$, where X is an outcome/observation or

*Corresponding author
Email addresses: `pier.bissiri@unimib.it` (Pier Giovanni Bissiri), `s.g.walker@kent.ac.uk` (Stephen G. Walker)

piece of information and $a \in \mathcal{A}$, then the best choice is that $\widehat{a}$ which minimizes $L(a, X)$. On the other hand, if there are a number of pieces of information, say $(X_1, \ldots, X_n)$, each of which contributes an additive loss $L(a, X_i)$ under action a, then the best choice now minimizes the cumulative loss

$$L(a, X_1, \ldots, X_n) = \sum_{i=1}^{n} L(a, X_i).$$

Such an additive style of cumulative loss would be appropriate when the (X_i) are independent pieces of information.

We are interested in the case when $\mathcal{A}$ is the space of probability distributions on $(0, 1)$. This occurs if our aim is to choose a probability distribution representing beliefs about a model parameter θ belonging to $(0, 1)$. In this framework, we allow $\pi(\cdot)$ to be the proposed representation of beliefs about θ in the case of no observations. The distribution π represents information, just as the Bernoulli observations $(X_1, \ldots, X_n)$ represent information. Hence, in the case $n = 0$, the loss function is $l_\pi(a, \pi)$. Maintaining the idea of cumulative loss, we now have

$$L(a, X_1, \ldots, X_n, \pi) = \sum_{i=1}^{n} L(a, X_i) + l_\pi(a, \pi). \tag{1}$$

To this point there is little justification required; we are merely writing down a general loss function in order to determine a probability distribution on $(0, 1)$, where the only assumption is that the losses are additive or cumulative. This seems relevant when the pieces of information are independent; that is, no one piece of information provides information about any of the others. To better indicate that $\mathcal{A}$ is a set of probability distributions, we will now replace a with ν.

The first and straightforward loss function to discuss is $l_\pi(\nu, \pi)$. So, $l_\pi(\nu, \pi)$ is the loss when ν is the probability measure correctly representing beliefs (and indeed with consistency it will end up providing correct beliefs) and π is the proposed probability measure representing beliefs at the outset. Therefore, $l_\pi(\nu, \pi)$ can be interpreted as a loss in information. It is reasonable to require ν to be absolutely continuous with respect to π. Indeed, the updated probability should be zero on every event whose prior π probability is zero. Therefore, $l_\pi(\nu, \pi)$ can be taken to be the g–divergence, introduced by Ali and Silvey (1966) and Csiszár (1967), i.e. $l_\pi(\nu, \pi) = \int g(\mathrm{d}\nu/\mathrm{d}\pi)\, \mathrm{d}\nu$, where g is a convex function such that $g(1) = 0$. Such a family of divergences is known to be a generalization of the Kullback–Leibler divergence, introduced by Kullback and Leibler (1951), which is obtained taking $g(x) = -\log(x)$. Bissiri and Walker (2009) establish that among the g–divergences, the Kullback–Leibler divergence is the only one which preserves a necessary coherence property whereby the solution at stage n serves as the prior for subsequent observations. Hence, $l_\pi(\nu, \pi)$, the loss in information in using π rather than ν is taken to be the Kullback–Leibler divergence.

Our aim now is to ascertain how π changes to ν, in the light of the information $(X_1, \ldots, X_n)$, for an apparent arbitrary loss function $L(\nu, X)$. Our

form for this seems obvious in the sense that if we select $L(\theta, X)$ then we can merely take $L(\nu, X) = \int l(\theta, X)\,\nu(\mathrm{d}\theta)$, since the ν represents beliefs in θ and so $L(\nu, X)$ is understood as expected loss. Surprisingly now, an obvious asymptotic requirement will pin down $l(\theta, X)$ precisely. The following can also be seen as providing an explicit answer to a suggestion in Walker (2006, Section 6) about a possible justification of the Bayesian paradigm through the loss (1) and asymptotic requirements. So, while loss functions are typically regarded as a subjective choice, an objective choice based on asymptotic properties for the ν provides justification for the Bayesian learning process. All the proofs are deferred to the Appendix.

2. Preliminaries

Denote by $(X_n)_{n\geq 1}$ the sequence of observations which are i.i.d. Bernoulli with parameter θ_0. Assume they are $0-1$ random variables on a probability space $(\Omega, \mathscr{F}, P_\theta)$ and $P_\theta(X_n = 1) = \theta$ for each $n \geq 1$. Denote by π the prior distribution for θ. So, π is a probability measure on the Borel subsets of $(0, 1)$. In the rest of the paper, it will be assumed that π is absolutely continuous with respect to the Lebesgue measure λ and that there is a version of $\mathrm{d}\pi/\mathrm{d}\lambda$ that is a continuous function.

The observations $X_1, X_2, \ldots$ are usually considered conditionally independent and identically distributed given θ. See for example Bernardo and Smith (1994). This makes possible to update the prior π by Bayes' theorem, obtaining the posterior distribution for θ

$$\pi^{(n)}(A) := \pi(A \mid X_1, \ldots, X_n) = \frac{\int_A \theta^{n\hat{\theta}_n}\,(1-\theta)^{n(1-\hat{\theta}_n)}\,\pi(\mathrm{d}\theta)}{\int_{(0,1)} \theta^{n\hat{\theta}_n}\,(1-\theta)^{n(1-\hat{\theta}_n)}\,\pi(\mathrm{d}\theta)},$$

where A is a Borel subset of $(0, 1)$ and $\hat{\theta}_n := \frac{1}{n}\sum_{i=1}^n X_i$.

Applying Bayes' theorem to obtain the posterior distribution, the observations are not considered independent, but conditionally independent given θ. Since we are assuming that the observations are independent, we are uncomfortable with the notion of a Bayesian model which artificially creates a dependence between the observations. However, as it will be soon clear, the posterior distribution also arises as the solution of a minimization problem, which does not require such an assumption of dependence for the observations.

Following Section 1, we consider the loss (1) taking $l(\theta, X) = -\ln(P_\theta(X_1 = X)/P_{\theta_0}(X_1 = X))$, the self–information loss function. So, the loss function (1) becomes:

$$\begin{aligned} L(\nu) &:= L(\nu, x_1, \ldots, x_n, \pi) \\ &= -\sum_{i=1}^n \int_{(0,1)} \ln(P_\theta(X_1 = x_i)/P_{\theta_0}(X_1 = x_i))\;\nu(\mathrm{d}\theta) \;+\; D(\nu, \pi) \end{aligned} \tag{2}$$

where $(x_1, \ldots, x_n)$ is a sample drawn from $(X_1, \ldots, X_n)$, ν is a probability measure on $(0, 1)$ absolutely continuous with respect to π, and D denotes the

Kullback–Leibler divergence (relative entropy), i.e.

$$D(Q_1, Q_2) = \int_S \ln\left(\frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}\right) \mathrm{d}Q_1,$$

where S is the support of Q_1, for any couple (Q_1, Q_2) of probability measures such that $Q_1 \ll Q_2$.

Notice that the first addendum in (2) depends on the sample and attains its minimum when $\nu = \delta_{\hat{\theta}_n}$, i.e. ν is degenerate at $\hat{\theta}_n$, while the second term takes into account only the prior belief about θ expressed by π. It is clear that the posterior $\pi^{(n)}$ minimizes the loss L, since

$$L(\nu) = D(\nu, \pi^{(n)}) - \ln\left(\int_{(0,1)} \prod_{i=1}^{n} P_\theta(X_1 = x_i)\, \pi(\mathrm{d}\theta)\right) + \sum_{i=1}^{n} \ln(P_{\theta_0}(X_1 = x_i)).$$

We could stop here since we have a justifiable loss function the solution of which is the Bayesian posterior distribution. However, we can work with a more general loss function and establish through asymptotic arguments that the $-\log$ loss is the best in some sense.

So, an obvious and general alternative for the loss function in (2) is

$$\begin{aligned} L_f(\nu) := & \sum_{i=1}^{n} \int_{(0,1)} f(P_\theta(X_1 = x_i))\ \nu(\mathrm{d}\theta) \\ & + \int_{(0,1)} \ln\left(\frac{\nu(\mathrm{d}\theta)}{\pi(\mathrm{d}\theta)}\right) \nu(\mathrm{d}\theta), \quad \nu \ll \pi, \end{aligned} \tag{3}$$

where the function $-\ln(\cdot)$ has been replaced by a function f from $(0,1)$ into the non–negative real line including $+\infty$. Clearly this is appropriate as $f(P_\theta(X_1 = x))$ is either $f(\theta)$ or $f(1-\theta)$, depending on whether x is 1 or 0. Denote by $\pi_f^{(n)}$ the probability measure that is absolutely continuous with respect to π with density

$$\frac{e^{-n\{\hat{\theta}_n f(\theta) + (1-\hat{\theta}_n) f(1-\theta)\}}}{\int_{(0,1)} e^{-n\{\hat{\theta}_n f(t) + (1-\hat{\theta}_n) f(1-t)\}}\, \mathrm{d}\pi(t)}.$$

The probability measure $\pi_f^{(n)}$ minimizes L_f since

$$L_f(\nu) = D(\nu, \pi_f^{(n)}) - \ln\left(\int_{(0,1)} e^{-n\{\hat{\theta}_n f(\theta) + (1-\hat{\theta}_n) f(1-\theta)\}}\, \pi(\mathrm{d}\theta)\right).$$

A referee has pointed out connections with Lagrange functions, $\pi_f^{(n)}$ and the unique minimization of L_f.

3. Theory

Our aim is to properly choose the function f within the class $C^1(0,1)$, apart from an additive constant. In fact, for any real constant c, $\pi_f^{(n)}(\cdot) = \pi_{f+c}^{(n)}(\cdot)$.

It will be shown which conditions on f are necessary and sufficient for the (strong) consistency of $\pi_f^{(n)}$. Next, a criterion will be defined that makes $f(\cdot) = -\ln(\cdot)$ the best choice and therefore the Bayesian posterior $\pi^{(n)}$ the best one.

3.1. Consistency

Assume that

$$\pi(\theta_0 - \varepsilon, \theta_0 + \varepsilon) > 0, \tag{4}$$

for every $\varepsilon > 0$. Since θ_0 is unknown, this means that only priors whose support is the unit interval will be considered. It will be convenient to express the posterior in the following form:

$$\pi_f^{(n)}(A) = \frac{\int_A \; e^{-n\, d(t,\hat{\theta}_n)}\, \mathrm{d}\pi(t)}{\int_{(0,1)} e^{-n\, d(t,\hat{\theta}_n)}\, \mathrm{d}\pi(t)}, \tag{5}$$

where

$$d(x,y) := \; y\,(f(x) - f(y)) \; + \; (1-y)\,(f(1-x) - f(1-y)) \tag{6}$$

for every (x, y) in $(0, 1)^2$.

Proposition 1. *Let f be a function of class $C^1(0,1)$ and let d be a function defined on $(0,1)^2$ by (6).*

Then the following facts are equivalent:

(i) For every θ_0, $\varepsilon > 0$, and every prior π satisfying (4),

$$\pi_f^{(n)}(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \longrightarrow 1, \quad n \to \infty, \qquad P_{\theta_0} - \; a.s. \tag{7}$$

(ii) For every (θ_1, θ_2) in $(0,1)^2$,

a) $d(\theta_1, \theta_2) \geq 0$

b) if $\theta_1 \neq \theta_2$ then $d(\theta_1, \theta_2) > 0$.

(iii) For every $0 < x < 1$,

a) $x\, f'(x) \; = \; (1-x)\, f'(1-x)$,

b) $f'(x) \; < \; 0$.

In the rest of the paper, it will be assumed that π is absolutely continuous with respect to the Lebesgue measure and that its density is continuous on (0, 1). The following proposition determines the rate of convergence of $\pi_f^{(n)}$. In its statement and in the rest of the paper, the minimum between two real numbers x and y will be denoted by $x \wedge y$.

Proposition 2. *If the hypotheses of Proposition 1 are satisfied, then, as $n \to \infty$,*

$$\pi_f^{(n)}((\theta_0 - \varepsilon, \theta_0 + \varepsilon)^c) \asymp \frac{1}{\sqrt{n}}\, e^{-n(\delta(\varepsilon))} \qquad P_{\theta_0} - \; a.s., \tag{8}$$

where $\delta(\varepsilon) := d(\theta_0 - \varepsilon, \theta_0) \wedge d(\theta_0 + \varepsilon, \theta_0)$.

3.2. The choice of the loss function.

Here we study the large sample property of the posterior, and this can be done by considering the posterior variance given $\hat{\theta}_n$.

Proposition 3. *If $V_f(\hat{\theta}_n)$ denotes the variance with respect to the distribution $\pi_f^{(n)}$ and f satisfies the conditions of Proposition 1, then*

$$\lim_{n\to\infty} nV_f(\hat{\theta}_n) \;=\; -\frac{\theta_0(1-\theta_0)}{\theta_0\, f'(\theta_0)}, \qquad P_{\theta_0} - a.s. \tag{9}$$

So, the limit depends obviously on θ_0 and it is clear how: the Fisher information is $I(\theta_0)^{-1} \propto \theta_0(1-\theta_0)$ and so for larger values of $I(\theta_0)$ we will have a faster rate of convergence since there is more information in the data for such θ_0. Moreover, the information tells us how the convergence depends on θ_0. This is amplified by the speed at which the $\hat{\theta}_n$ converges to θ_0, and is proportional to $\theta_0(1-\theta_0)$. So, the $\theta_0(1-\theta_0)$ term in the limit of (9) is taking account of the value of θ_0. The other term should therefore not depend on θ_0.

The reason for this is quite simple: if the $\theta_0\, f'(\theta_0)$ does depend on θ_0 then we should be able to modify f so that all θ_0 obtain the largest value of $|\theta f'(\theta)|$. Hence, the optimal f must indeed make this a constant and so we must take $-xf'(x) = M$ for some constant $M > 0$. Hence we have $f(x) = -M \ln x$.

Hence, we now just need to ascertain the reason why we should make $M = 1$; since we have established that we must have $f(x) = -M \ln x$ and the Bayesian learning rule is obtained precisely with $M = 1$. Suppose the choice $\theta = \theta_1$ is chosen stubbornly so that $\pi(\theta) = \delta_{\theta_1}(\theta)$. Hence, since δ_{θ_1} will always represent beliefs, according to definitions in Section 2,

$$L(\delta_{\theta_1}) = M \sum_{i=1}^{n} \ln \left\{ \frac{P_{\theta_0}(X_1 = x_i)}{P_{\theta_1}(X_1 = x_i)} \right\}$$

and so our expected loss for n observations is

$$\bar{L}(\delta_{\theta_1}) = nMD(P_{\theta_0}, P_{\theta_1}).$$

We can understand that our loss up to a sample of size n when fixing θ_1; it is predicting with the wrong measure, i.e. P_{θ_1} instead of P_{θ_0} on n occasions. So our loss is $nD(P_{\theta_0}, P_{\theta_1})$, being consistent with using $D(\nu, \pi)$ in Section 1, and hence we must fix $M = 1$.

4. Discussion

We have constructed a loss function for selecting an updated belief probability measure on $(0, 1)$ in the light of i.i.d. Bernoulli random variables. Having started out with a general form, the precise function can be pinned down by appealing to some necessary asymptotic properties. The consequence is that the Bayesian learning machine, in the Bernoulli case at least, can be understood via

notions of loss functions and asymptotics and while retaining the correct notion of an i.i.d. sample.

We believe that the ideas in this paper can be extended to the more general case; in the first instance for parametric models $f(x;\theta)$ and subsequently for nonparametric models $f(x)$, $f \in \mathcal{F}$, where now the decision space consists of probability measures on $\mathcal{F}$.

Appendix

In order to prove Proposition 1, the following lemma will be useful.

Lemma 4. *Let f be a function of class $C^1(0,\,1)$ such that*

a) $x\,f'(x) \;=\; (1-x)\,f'(1-x)$,

b) $f'(x) \;<\; 0$,

for every $0 < x < 1$. Moreover, fix $0 < \theta < 1$ and define

$$\varphi(x) \;=\; \theta\,f(x) \;+\; (1-\theta)\,f(1-x), \tag{10}$$

for every $0 < x < 1$.

Hence, φ is a function of class $C^1(0,\,1)$ such that

$$\varphi'(x) \;=\; \frac{\theta - x}{1-x}\,f'(x). \tag{11}$$

Moreover, φ has the second derivative at θ, which is equal to

$$\varphi''(\theta) \;=\; -\frac{f'(\theta)}{1-\theta}. \tag{12}$$

Proof. By (10),

$$\varphi'(x) \;=\; \theta f'(x) \;-\; (1-\theta)\,f'(1-x). \tag{13}$$

A combination of (13) with (a) yields (11). Since f' is a continuous function, (11) entails

$$\lim_{x\to\theta} \frac{\varphi'(x) - \varphi'(\theta)}{x-\theta} \;=\; -\frac{f'(\theta)}{1-\theta},$$

and (12) is proved. $\square$

Proof of Proposition 1. Let $A^c := (0,1) \setminus A$ denote the complement of subset A of $(0,1)$.

To begin with, notice that by (5)

$$\pi_f^{(n)}(\theta_0 - \varepsilon, \theta_0 + \varepsilon) = \left(1 + \frac{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c} e^{-n\,d(\theta,\hat{\theta}_n)}\pi(\mathrm{d}\theta)}{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)} e^{-n\,d(\theta,\hat{\theta}_n)}\pi(\mathrm{d}\theta)}\right)^{-1},$$

and therefore (7) is tantamount to

$$\lim_{n\to\infty} \frac{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c} e^{-n\, d(\theta,\hat{\theta}_n)}\pi(\mathrm{d}\theta)}{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)} e^{-n\, d(\theta,\hat{\theta}_n)}\pi(\mathrm{d}\theta)} = 0 \qquad P_{\theta_0} - \text{a.s.} \tag{14}$$

Let us prove that (ii) is necessary for (i). To this aim, assume that (i) is true and fix $0 < \theta_0 < 1$ and a probability measure π satisfying (4). Hence, by virtue of (i), π must satisfy (7) as well.

Since f is continuous, the function $d(\cdot,\theta_1)$ is continuous as well for every θ_1 in $(0,1)$. In particular, for every θ_1, it is continuous at θ_1 where its value is zero; i.e. for every $k > 0$ there exists some $\varepsilon > 0$ such that $\inf_{\theta:\,|\theta-\theta_1|\,<\,2\varepsilon} d(\theta,\theta_1) > -2k$. Moreover, by the strong law of large numbers, $|\hat{\theta}_n - \theta_0| < \varepsilon$ for sufficiently large n P_{θ_0}-a.s., and therefore

$$\inf_{\theta:\,|\theta-\theta_0|\,<\,\varepsilon} d(\theta,\hat{\theta}_n) \geq \inf_{\theta:\,|\theta-\hat{\theta}_n|\,<\,2\varepsilon} d(\theta,\hat{\theta}_n) > -2k,$$

for sufficiently large n, P_{θ_0}-a.s., for every $k > 0$ and some $\varepsilon > 0$. Hence, by the dominated convergence theorem,

$$\begin{aligned} \lim_{n\to\infty} e^{-2nk} & \int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)} e^{-n\, d(\theta,\hat{\theta}_n)}\pi(\mathrm{d}\theta) \\ & = \int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)} \lim_{n\to\infty} e^{-n\,(d(\theta,\hat{\theta}_n)+2k)}\pi(\mathrm{d}\theta) = 0 \qquad P_{\theta_0} - \text{a.s.}, \end{aligned} \tag{15}$$

for every $k > 0$ and for some $\varepsilon > 0$.

Now assume that there is some M such that $d(\theta,\theta_0) \leq M$ for every θ belonging to some set C such that $\pi(C \setminus \{\theta_0\}) > 0$. Take $\varepsilon \geq 0$ small enough so that $\pi((\theta_0-\varepsilon,\theta_0+\varepsilon)^c \cap C) > 0$. Notice that, by the strong law of large numbers and by continuity of $d(\theta,\cdot)$, for every $\theta \in (0,1)$, $d(\theta,\hat{\theta}_n) - d(\theta,\theta_0) < M$ for sufficiently large n, P_{θ_0}-a.s. Hence, for every $\theta \in C$, $d(\theta,\hat{\theta}_n) < 2M$ for sufficiently large n, P_{θ_0}-a.s. and, by Fatou's lemma,

$$\begin{aligned} \liminf_{n\to\infty} e^{2nM} & \int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c} e^{-n\, d(\theta,\hat{\theta}_n)}\pi(\mathrm{d}\theta) \\ & \geq \int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c\cap C} \liminf_{n\to\infty} e^{n\,(2M-d(\theta,\hat{\theta}_n))}\,\pi(\mathrm{d}\theta) = \infty, \qquad P_{\theta_0} - \text{a.s.} \end{aligned} \tag{16}$$

So, combining (15) and (16), one notices that

$$\frac{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c} e^{-n\, d(\theta,\hat{\theta}_n)}\pi(\mathrm{d}\theta)}{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)} e^{-n\, d(\theta,\hat{\theta}_n)}\pi(\mathrm{d}\theta)} \geq e^{-2n(M+k)} \tag{17}$$

holds for sufficiently large n, P_{θ_0}-a.s., for every $k > 0$ and some $\varepsilon > 0$. Since (14) holds true for every $\varepsilon \geq 0$, then $M \geq -k$ for any real positive number k,

i.e. $M \geq 0$. So, $M \geq 0$ whenever $d(\theta, \theta_0) \leq M$ with positive π-probability. Therefore, $d(\theta, \theta_0) \geq 0$, π-a.s. Hence,

$$
\begin{aligned}
&\lim_{n\to\infty} \int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)} e^{-n\, d(\theta,\hat{\theta}_n)} \pi(\mathrm{d}\theta) \\
&\quad = \int_{\{\theta\in(\theta_0-\varepsilon,\theta_0+\varepsilon):\, d(\theta,\theta_0)\,>\,0\}} \lim_{n\to\infty} e^{-n\, d(\theta,\hat{\theta}_n)} \pi(\mathrm{d}\theta) \\
&\qquad + \lim_{n\to\infty} \int_{\{\theta\in(\theta_0-\varepsilon,\theta_0+\varepsilon):\, d(\theta,\theta_0)\,=\,0\}} e^{-n\, d(\theta,\hat{\theta}_n)} \pi(\mathrm{d}\theta) \\
&\quad = \pi\{\theta \in (\theta_0-\varepsilon, \theta_0+\varepsilon) :\, d(\theta,\theta_0) = 0\} \\
&\quad \leq \pi(\theta_0-\varepsilon, \theta_0+\varepsilon)\} \qquad\qquad P_{\theta_0} - \text{a.s.},
\end{aligned}
\tag{18}
$$

holds true P_{θ_0}-a.s., by dominated convergence theorem. Assume there is some M such that $d(\theta, \theta_0) \leq M$, for every θ belonging to some set D with positive π-probability and such that $\theta_0 \notin D$. Hence, combining (18) and (16), one notices that

$$
\frac{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c} e^{-n\, d(\theta,\hat{\theta}_n)} \pi(\mathrm{d}\theta)}{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)} e^{-n\, d(\theta,\hat{\theta}_n)} \pi(\mathrm{d}\theta)} \geq \pi(\theta_0-\varepsilon, \theta_0+\varepsilon)\, e^{-2nM} \tag{19}
$$

holds true for sufficiently large n, P_{θ_0}-a.s., for sufficiently small $\varepsilon > 0$. Since (14) holds true for every $\varepsilon \geq 0$ together with (4), then M must be positive. So, $M > 0$ whenever $d(\theta, \theta_0) \leq M$ for every $\theta \neq \theta_0$ with positive π-probability. Therefore, $d(\theta, \theta_0) > 0$, for every $\theta \neq \theta_0$ π-a.s. Since this is true for every θ_0 and every π whose support is the unit interval, (ii.b) must hold. Notice that (ii.b) trivially entails (ii.a) since $d(\theta, \theta) = 0$ for every θ by definition of d.

At this point, it will be proved that (ii) implies (iii). To this aim, define $\varphi(x) := d(x, \theta)$ for a fixed $0 < \theta < 1$. Since $d(\theta, \theta) = 0$, condition (ii) is tantamount to say that the function φ has an absolute minimum at $x = \theta$ for any θ. Therefore, if (ii) is in force, $\varphi'(\theta) = 0$ must be true for every θ in the unit interval and condition (iii.a) follows.

By Lemma 4, (iii.a) entails

$$
\varphi'(x) \; = \; \frac{\theta - x}{1 - x}\, f'(x), \tag{20}
$$

where φ' is a continuous function since f' is so. Since θ is an absolute minimum point for φ, there is $\delta > 0$ such that $\varphi'(x) > 0$ if $\theta < x \leq \theta + \delta$ and $\varphi'(x) < 0$ if $\theta - \delta \leq x < \theta$. This is tantamount to say that $f'(x) < 0$ for every x in $(\theta - \delta, \theta) \cup (\theta, \theta + \delta)$ for some δ. Since this must hold for every θ, condition (iii.b) follows.

Finally, it will be shown that (iii) is sufficient for (i). To this aim, notice that if (iii) holds then (20) is also in force by Lemma 4 and therefore $d(\cdot, \theta)$ is (strictly) decreasing on $(0, \theta)$ and (strictly) increasing on $(\theta, 1)$, for every θ. Therefore, for every $\varepsilon > 0$ and every $0 < \theta_1 < 1$, $d(\theta, \theta_1) > \delta(\varepsilon, \theta_1)/2$ if $|\theta - \theta_1| > \varepsilon$ and $\delta(\varepsilon, \theta_1)$ denotes $d(\theta_1 - \varepsilon, \theta_1) \wedge d(\theta_1 + \varepsilon, \theta_1)$. Applying dominated convergence

theorem, this entails that

$$\begin{aligned}\lim_{n\to\infty} e^{n\delta(\varepsilon,\hat\theta_n)/2} \int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c} & e^{-n\,d(\theta,\hat\theta_n)}\,\pi(\mathrm{d}\theta)\\ &= \int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c} \lim_{n\to\infty} e^{-n\,(d(\theta,\hat\theta_n)-\delta(\varepsilon,\hat\theta_n)/2)}\,\pi(\mathrm{d}\theta) = 0.\end{aligned} \tag{21}$$

By continuity of $d(\cdot,\hat\theta_n)$ at $\hat\theta_n$, for every $\eta > 0$ there exists γ such that $d(\theta,\hat\theta_n) < \eta$ if $|\theta-\hat\theta_n| < 2\gamma$ and by the strong law of large numbers $|\hat\theta_n - \theta_0| < \gamma$ for sufficiently large n, P_{θ_0}-a.s. Therefore $d(\theta,\hat\theta_n) < \eta$ if $|\theta-\theta_0| < \gamma$ for sufficiently large n, $P_{\theta-0}$-a.s., and by Fatou's lemma,

$$\begin{aligned}\liminf_{n\to\infty} e^{n\eta} \int_{(0,1)} & e^{-n\,d(\theta,\hat\theta_n)}\,\pi(\mathrm{d}\theta)\\ &\geq \int_{(\theta_0-\gamma,\theta_0+\gamma)} \liminf_{n\to\infty} e^{n\,(\eta-d(\theta,\hat\theta_n))}\,\pi(\mathrm{d}\theta) = \infty,\end{aligned} \tag{22}$$

for every $\eta > 0$ and some $\gamma > 0$, $P_{\theta-0}$-a.s..

Combining (21) and (22), one obtains that

$$\lim_{n\to\infty} e^{n(\delta(\varepsilon,\hat\theta_n)/2-\eta)} \frac{\int_{(\theta_0-\varepsilon,\theta_0+\varepsilon)^c} e^{-n\,d(\theta,\hat\theta_n)}\,\pi(\mathrm{d}\theta)}{\int_{(0,1)} e^{-n\,d(\theta,\hat\theta_n)}\,\pi(\mathrm{d}\theta)} = 0,$$

for every $\eta,\varepsilon > 0$. Taking $\eta < \delta(\varepsilon,\theta_0)/2$, (i) follows.

□

By the strong law of large numbers, there exists a Borelian subset B of $\{0,1\}^\infty$ with P_{θ_0}-probability one such that $\hat\theta_n(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^n x_i$ converges to θ_0 for all sequences $(x_n)_{n\geq 1}$ belonging to B. In the rest of this appendix, $\hat\theta_n$ will stand for $\hat\theta_n(x_1,\ldots,x_n)$ and we shall always assume that $(x_n)_{n\geq 1}$ belongs to B.

In order to prove Proposition 2 and Proposition 3, the following lemmas are useful.

Lemma 5. *If $(g_n)_{n\geq 0}$ is a sequence of non-negative functions on $(0,1)$ dominated by an integrable function, then for every $\delta > 0$ there are η_1, $c_0 > 0$ such that*

$$\int_{(\hat\theta_n-\delta,\,\hat\theta_n+\delta)^c} e^{-nd(t,\hat\theta_n)}\,g_n(t)\,\mathrm{d}t \leq c_0\,e^{-n\eta_1} \tag{23}$$

for sufficiently large n.

Moreover, if $(c_n)_{n\geq 1}$ is a sequence converging to a positive number, $(g_n^*)_{n\geq 1}$ is a sequence of integrable functions on $\mathbb{R}$, and

$$\int_{(-\infty,\,\infty)} e^{-c_n(t-\hat\theta_n)^2/2}\,g_n^*(t)\,\mathrm{d}s \,<\, c, \tag{24}$$

for some real constant c and for sufficiently large n, then

$$\int_{(\hat{\theta}_n-\delta,\,\hat{\theta}_n+\delta)^c} e^{-n\,c_n(t-\hat{\theta}_n)^2/2}\, g_n^*(t-\hat{\theta}_n)\,\mathrm{d}t \le k\, e^{-n\eta_2}, \tag{25}$$

for some $k,\, \eta_2 > 0$ and for sufficiently large n.

In the rest of the paper, the maximum between two real numbers x and y will be denoted by $x \vee y$.

Proof. Let χ_n be a nonnegative and differentiable function on $(a,\, b)$ $(-\infty \le a, b \le \infty)$ with an unique absolute minimum at $\hat{\theta}_n$ and such that

$$\begin{aligned} &\chi_n'(t) < 0 \qquad \text{if} \quad a < t < \hat{\theta}_n \\ &\chi_n'(t) > 0 \qquad \text{if} \quad \hat{\theta}_n < t < b. \end{aligned} \tag{26}$$

Hence, if $\delta > 0$ and $|t - \hat{\theta}_n| > \delta$ then $\chi_n(t) > \eta_n(\delta)$, where $\eta_n(\delta) := \chi_n(\hat{\theta}_n - \delta) \vee \chi_n(\hat{\theta}_n + \delta)$. Notice that $-n\,\chi_n(t) \le -(n-1)\eta_n(\delta) - \chi_n(t)$ if $|t - \hat{\theta}_n| > \delta$ and $n > 1$. Therefore, given some sequence of measures $(\mu_n)_n$ on the Borelian subsets of $(a,\, b)$,

$$\int_{\{|t|\,>\,\delta\}} e^{-n\chi_n(t)}\mu_n(\mathrm{d}t) < e^{-(n-1)\eta_n(\delta)} \int_{(a,\,b)} e^{-\,\chi_n(t)}\mu_n(\mathrm{d}t). \tag{27}$$

Taking $\chi_n(t) = d(t, \hat{\theta}_n)$, $a = 0$, $b = 1$, (26) holds true. Moreover, the integral

$$\int_{(a,\,b)} e^{-\,\chi_n(t)}\mu_n(\mathrm{d}t) \tag{28}$$

is less than a constant, if $\mathrm{d}\mu_n/\mathrm{d}\lambda = g_n$ and λ is the Lebesgue measure. In fact, χ_n is nonnegative and g_n dominated. Since $\eta_n(\delta) = d(\hat{\theta}_n - \delta, \hat{\theta}_n) \vee d(\hat{\theta}_n + \delta, \hat{\theta}_n)$ converges to a positive constant by the strong law of large numbers, (27) yields (23).

If $\chi_n(t) = c_n(t - \hat{\theta}_n)^2/2$, $a = -\infty$, $b = \infty$, then (26) is satisfied. Moreover, if $\mathrm{d}\mu_n/\mathrm{d}\lambda(t) = g_n^*(t - \hat{\theta}_n)$, then the integral (28) turns out to be equal to the integral in (24). Therefore, (25) follows from (27). □

Lemma 6. *Let $(g_n)_{n\ge1}$ and $(g_n^*)_{n\ge1}$ be two sequences of nonnegative, continuous and integrable functions defined on $(0,\, 1)$ and $\mathbb{R}$, respectively, and such that $g_n(t) \sim g_n^*(t)$ as $t \to \hat{\theta}_n$, for every $n \ge 1$.*

Let $d_n(t)$ stand for $d(t, \hat{\theta}_n)$, and denote:

$$I_n := \int_{(0,1)} e^{-nd_n(t)}\, g_n(t)\,\mathrm{d}t, \tag{29}$$

$$I_n(x) := \int_{(-\infty,\,\infty)} e^{-n(d_n''(\hat{\theta}_n)-x)(t-\hat{\theta}_n)^2/2}\, g_n^*(t)\,\mathrm{d}t, \tag{30}$$

Assume that

$$\lim_{n\to\infty} e^{-c\,n}/(I_n(x) - I_n(y)) = 0, \tag{31}$$

for every $c > 0$, and every x, y belonging to some neighborhood of zero. Moreover, let (25) hold with $c_n = d''_n(\hat{\theta}_n)$, for some positive constants k, η_2.

Therefore, $I_n \sim I_n(0)$ as $n \to \infty$.

Proof. This proof will be based on the Laplace method. See, for instance, de Bruijn (1981, pp. 63–65). His results do not precisely fit our needs and therefore we have to prove this lemma starting from scratch.

Recall that $d_n(\hat{\theta}_n) = 0$. Moreover, by hypothesis, d_n has a unique minimum at $\hat{\theta}_n$ so that $d'_n(\hat{\theta}_n) = 0$. By Taylor's theorem, for each $n \geq 1$ and each $\varepsilon > 0$ there exists $\delta_n > 0$ such that if $|t - \theta_0| < \delta_n$ then

$$|d_n(t) - \frac{1}{2}\, d''_n(\hat{\theta}_n)\,(t - \hat{\theta}_n)^2| \,<\, \varepsilon\,(t - \hat{\theta}_n)^2. \tag{32}$$

It will be useful to observe that δ_n can be taken constant for sufficiently large n. In order to show this fact, define

$$\psi_n(t) \,:=\, d_n(t) \,-\, \frac{1}{2} d''_n(\hat{\theta}_n)(t - \hat{\theta}_n)^2,$$

so that $\psi_n(\hat{\theta}_n) = \psi'_n(\hat{\theta}_n) = \psi''_n(\hat{\theta}_n) = 0$. By (11) and (12),

$$\begin{aligned} \psi'_n(t) \,&:=\, d'_n(t) \,-\, d''_n(\hat{\theta}_n)(t - \hat{\theta}_n) \\ &=\, (\hat{\theta}_n - t)\left(\frac{f'(t)}{1-t} - \frac{f'(\hat{\theta}_n)}{1 - \hat{\theta}_n}\right). \end{aligned} \tag{33}$$

Recall that $0 < \theta_0 < 1$ and fix $0 < \gamma < (1 - \theta_0) \wedge \theta_0$. By hypothesis, the function $f'(t)/(1-t)$ is continuous over the compact set $[\theta_0 - \gamma, \theta_0 + \gamma]$ and therefore is uniformly continuous over that interval. Moreover, recall that by the strong law of large numbers, $\hat{\theta}_n$ belongs to $[\theta_0 - \gamma, \theta_0 + \gamma]$ if $n \geq N$ for some N. Hence, by (33) for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|\psi'_n(t)/(t - \hat{\theta}_n)| = |f'(t)/(1-t) - f'(\hat{\theta}_n)/(1 - \hat{\theta}_n)| < \varepsilon \tag{34}$$

if $|t - \hat{\theta}_n| \,<\, \delta$ and $n \geq N$. By Lagrange's mean value theorem,

$$\psi_n(t) \,=\, \psi_n(t) - \psi_n(\hat{\theta}_n) \,=\, (t - \hat{\theta}_n)\psi'_n(s) \tag{35}$$

for some s between t and $\hat{\theta}_n$. Combining (34) with (35), one obtains

$$|\psi_n(t)| \,=\, |(t - \hat{\theta}_n)\,\psi'_n(s)| \,\leq\, \varepsilon|t - \hat{\theta}_n||s - \hat{\theta}_n|.$$

Since $|s - \hat{\theta}_n| \leq |t - \hat{\theta}_n|$, (32) holds true for every $t \in (\theta_0 - \delta, \theta_0 + \delta)$ and every $n \geq N$.

For every $n \geq 1$, $g_n(t) \sim g^*_n(t)$ as $t \to \hat{\theta}_n$, by hypothesis. Hence, the function

$$t \longrightarrow \mathbb{I}_{\{\hat{\theta}_n\}^c}(t)\, g_n(t)/g^*_n(t) \,+\, \mathbb{I}_{\{\hat{\theta}_n\}}(t)$$

is continuous on the compact set $[\theta_0-\gamma,\theta_0+\gamma]$ and therefore uniformly continuous on that set. For this reason, for each $\varepsilon>0$ there is $\delta>0$ such that

$$g_n^*(t)(1-\varepsilon) \;\le\; g_n(t) \;\le\; g_n^*(t)(1+\varepsilon) \tag{36}$$

if $|t-\hat\theta_n|<\delta$ and n is sufficiently large.

At this stage, fix ε belonging to $\big(\,0,\,-f'(\theta_0)/\{4(1-\theta_0)\}\,\big)$ so that

$$\varepsilon < -f'(\hat\theta_n)/\{3(1-\hat\theta_n)\} = d_n''(\hat\theta_n)/3 \tag{37}$$

for $n\ge M$ and some $M\ge N$. Moreover, take $\delta>0$ small enough so that (32) and (36) are both satisfied.

Decompose the integral I_n defined by (29) in the following way:

$$I_n = \int_{(\hat\theta_n-\delta,\,\hat\theta_n+\delta)} e^{-nd_n(t)}\,g_n(t)\,\mathrm{d}t \;+\; \int_{(\hat\theta_n-\delta,\,\hat\theta_n+\delta)^c} e^{-nd_n(t)}\,g_n(t)\,\mathrm{d}t.$$

The first term can be bounded by (32), the second one by (23), obtaining

$$\begin{aligned} I_n &\ge \int_{(\hat\theta_n-\delta,\,\hat\theta_n+\delta)} e^{-n(d_n''(\hat\theta_n)+2\varepsilon)(t-\hat\theta_n)^2/2}\,g_n(t)\,\mathrm{d}t \\ I_n &\le \int_{(\hat\theta_n-\delta,\,\hat\theta_n+\delta)} e^{-n(d_n''(\hat\theta_n)-2\varepsilon)(t-\hat\theta_n)^2/2}\,g_n(t)\,\mathrm{d}t + c_0e^{-n\eta_1}, \end{aligned}$$

which in virtue of (36) becomes

$$I_n \ge (1-\varepsilon)\int_{(\hat\theta_n-\delta,\,\hat\theta_n+\delta)} e^{-n(d_n''(\hat\theta_n)+2\varepsilon)(t-\hat\theta_n)^2/2}\,g_n^*(t)\,\mathrm{d}t \tag{38}$$

$$I_n \le\; (1+\varepsilon)\int_{(\hat\theta_n-\delta,\,\hat\theta_n+\delta)} e^{-n(d_n''(\hat\theta_n)-2\varepsilon)(t-\hat\theta_n)^2/2}\,g_n^*(t)\,\mathrm{d}t + c_0e^{-n\eta_1}. \tag{39}$$

By hypothesis, (25) holds true with $c_n = d_n''(\hat\theta_n)$, for some positive constants k,η_2. Therefore, (38) becomes

$$I_n \ge (1-\varepsilon)\int_{(-\infty,\,\infty)} e^{-n(d_n''(\hat\theta_n)+2\varepsilon)(t-\hat\theta_n)^2/2}\,g_n^*(t)\,\mathrm{d}t - (1-\varepsilon)k\,e^{-n\eta_2}. \tag{40}$$

Recalling (30), the combination of (40) and (39) yields

$$(1-\varepsilon)\,I_n(-2\varepsilon) \,-\, (1-\varepsilon)ke^{-n\eta_2} \le I_n \le (1+\varepsilon)I_n(2\varepsilon) \,+\, c_0\,e^{-n\eta_1} \tag{41}$$

for n sufficiently large. By (31), if n is sufficiently large, then

$$\begin{aligned} e^{-\eta_1 n} &\;<\; (1+\varepsilon)(I_n(3\varepsilon)-I_n(2\varepsilon))/c_0 \\ e^{-\eta_2 n} &\;<\; (I_n(-2\varepsilon)-I_n(-3\varepsilon))/k, \end{aligned}$$

being $I_n(x)$ an increasing function of x. Therefore, by (41),

$$(1-\varepsilon)\,I_n(-3\varepsilon) \;\le\; I_n \;\le\; (1+\varepsilon)\,I_n(3\varepsilon) \tag{42}$$

holds true for sufficiently large n. The number ε being arbitrary, it follows that $I_n \sim I_n(0)$ as $n\to\infty$

□

Lemma 7. *Let $d_n(t)$ stand for $d(t,\,\hat\theta_n)$. If the hypotheses and the conditions of Proposition 1 hold true, p is an integrable, nonnegative and continuous function on $(0,\,1)$ and $0\,<\,a\,<\,\theta_0\,<\,b\,<\,1$, then*

$$\int_{(0,\;1)} p(t)\, e^{-n\, d_n(t)}\, \mathrm{d}t \sim \frac{\sqrt{2\pi}\, p(\hat\theta_n)}{\sqrt{n\, d_n''(\hat\theta_n)}}, \tag{43}$$

$$\int_{(0,1)} p(t)\, e^{-n\, d_n(t)}\, (t-\hat\theta_n)^2\, \mathrm{d}t \sim \sqrt{2\pi}\, p(\hat\theta_n)\, \{n d_n''(\hat\theta_n)\}^{-3/2} \tag{44}$$

$$\int_{(0,1)} p(t)\, e^{-n\, d_n(t)}\, t^2\, \mathrm{d}t \sim \sqrt{2\pi}\, p(\hat\theta_n)\, \{n\, d_n''(\hat\theta_n)\}^{-1/2}\, (\hat\theta_n^2 + \{n\, d_n''(\hat\theta_n)\}^{-1}), \tag{45}$$

$$\int_{(b,\;1)} p(s)\, e^{-n\, d_n(s)}\, \mathrm{d}s \sim \frac{1}{n}\, \frac{p(b)}{d_n'(b)}\, e^{-n\, d_n(b)}, \tag{46}$$

$$\int_{(0,\;a)} p(s)\, e^{-n\, d_n(s)}\, \mathrm{d}s \sim -\frac{1}{n}\, \frac{p(a)}{d_n'(a)}\, e^{-n\, d_n(a)}, \tag{47}$$

as $n \to \infty$.

Proof. Lemma 6 will be applied to prove (43), (44) and (45). Three cases will be considered:

Case A) $g_n(t) = p(t), \quad g_n^*(t) = p(\hat\theta_n);$

Case B) $g_n(t) = p(t)(t-\hat\theta_n)^2, \quad g_n^*(t) = p(\hat\theta_n)(t-\hat\theta_n)^2;$

Case C) $g_n(t) = p(t)t^2, \quad g_n^*(t) = p(\hat\theta_n)t^2;$

Notice that the integral $I_n(x)$ defined by (30) is finite if $x\,<-\,f'(\theta_0)/\{2(1-\theta_0)\}$ and n is sufficiently large. In fact, by the strong law of large numbers, this entails that $x\,<-\,f'(\hat\theta_n)/(1-\hat\theta_n)$ for sufficiently large n, and therefore, by (12) in Lemma 4, $d_n''(\hat\theta_n) - x$ is positive.

If $x\,<-\,f'(\theta_0)/\{2(1-\theta_0)\}$, then

$$I_n(x) = \sqrt{\frac{2\pi}{n(d_n''(\hat\theta_n) - x)}}\; \mathbb{E}(g_n^*(W_n)),$$

where W_n is a Gaussian random variable with mean $\hat\theta_n$ and variance $\{n(d''(\hat\theta_n) - x)\}^{-1}$. Therefore,

Case A) $I_n(x) \,=\, \sqrt{2\pi}\, p(\hat\theta_n)\, \{n(d_n''(\hat\theta_n) - x)\}^{-1/2},$

Case B) $I_n(x) \,=\, \sqrt{2\pi}\, p(\hat\theta_n)\, \{n(d_n''(\hat\theta_n) - x)\}^{-3/2},$

Case C) $I_n(x) \,=\, \sqrt{2\pi}\, p(\hat\theta_n)\, \{n(d_n''(\hat\theta_n) - x)\}^{-1/2}\, (\hat\theta_n^2 + \{n(d''(\hat\theta_n) - x)\}^{-1}),$

if $x < -f'(\theta_0)/\{2(1-\theta_0)\}$ and n is sufficiently large.

In all three cases, (31) holds true if $c > 0$, $|x|, |y| < -f'(\theta_0)/\{2(1-\theta_0)\}$. Moreover, the integral in (24) with $c_n = n\, d_n''(\hat{\theta}_n)$ is equal to

Case A) $\sqrt{2\pi}\, p(\hat{\theta}_n)\, \{n\, d_n''(\hat{\theta}_n)\}^{-1/2}$,

Case B) $p(\hat{\theta}_n)\, \sqrt{2\pi}\, \{n\, d_n''(\hat{\theta}_n)\}^{-3/2}$,

Case C) $\sqrt{2\pi}\, p(\hat{\theta}_n)\, \{n\, d_n''(\hat{\theta}_n)\}^{-1/2}\, (\hat{\theta}_n^2 + \{n\, d_n''(\hat{\theta}_n)\}^{-1})$,

which converge to zero by continuity of d_n'' and p, and by the strong law of large numbers. Therefore, (24) is satisfied in all three cases if n is sufficiently large. This allows us to apply Lemma 5 and to obtain that (25) holds true for some positive constants k, η_2.

Since (31) and (25) hold for some positive constants k, η_2, Lemma 6 can be applied and (43), (44) and (45) are proved.

At this stage, our aim is to prove (46) and (47). The Laplace method will be used again. By continuity of the function p, for each $\varepsilon > 0$ there is $\delta > 0$ such that

$$p(b)(1-\varepsilon) \le p(t) \le p(b)(1+\varepsilon) \tag{48}$$

if $b \le t < b + \delta$.

Since f is continuous, the functions

$$\begin{aligned} &t \longrightarrow f(t) - f(b) - (t-b)f'(b) \\ &t \longrightarrow f(1-t) - f(1-b) - (b-t)f'(1-b) \end{aligned}$$

are continuous at b and at $1-b$. Therefore, for a given $\varepsilon > 0$ we can fix $\delta > 0$ such that

$$\begin{aligned} &\left| d(t,\, s) - d(b,\, s) - (t-b)\, \frac{\partial d(x,\, s)}{\partial x}\bigg|_{x=b} \right| \\ &\quad \le s\; |(f(t) - f(b) - (t-b)f'(b))| \\ &\qquad + (1-s)\, |f(1-t) - f(1-b) - (b-t)f'(1-b)| < \varepsilon \end{aligned}$$

holds for every $t \in (b,\, b+\delta)$ for and every $s \in (0,1)$.

Hence, for a given $\varepsilon > 0$, we can fix $\delta > 0$ such that

$$|d_n(t) - d_n(b) - (t-b)\, d_n'(b)| \le \varepsilon \tag{49}$$

holds true together with (48) for every $t \in (b,\, b+\delta)$.

Denote

$$J_n := \int_{(b,\,1)} e^{-n d_n(t)} p(t) \mathrm{d}t.$$

Since d_n is increasing on $(b+\delta,\, 1) \subset (\theta_0,\, 1)$,

$$\begin{aligned} J_n &= \int_{(b,\, b+\delta)} e^{-n d_n(t)} p(t) \mathrm{d}t + \int_{(b+\delta,\, 1)} e^{-n d_n(t)} p(t) \mathrm{d}t \\ &\le \int_{(b,\, b+\delta)} e^{-n d_n(t)} p(t) \mathrm{d}t + k\, e^{-n d_n(b+\delta)} \end{aligned} \tag{50}$$

where $k = \int_{(b+\delta,\,1)} p(t)\mathrm{d}t$. Therefore, we can write

$$\int_{(b,\,b+\delta)} e^{-nd_n(t)}p(t)\mathrm{d}t \le J_n \le \int_{(b,\,b+\delta)} e^{-nd_n(t)}p(t)\mathrm{d}t + k\,e^{-nd_n(b+\delta)},$$

which yields, by (48) and (49), for sufficiently large n,

$$\begin{aligned} J_n \;&\ge\; (p(b)-\varepsilon)\,e^{-nd_n(b)} \int_{(b,\,b+\delta)} e^{-n(d_n'(b)+\varepsilon)(t-b)}\mathrm{d}t \\ J_n \;&\le\; (p(b)+\varepsilon)\,e^{-nd_n(b)} \int_{(b,\,b+\delta)} e^{-n(d_n'(b)-\varepsilon)(t-b)}\mathrm{d}t + k\,e^{-nd_n(b+\delta)}, \end{aligned}$$

that is

$$(p(b)-\varepsilon)\,\bar{J}_n(-\varepsilon) \le J_n \le (p(b)+\varepsilon)\,\bar{J}_n(\varepsilon) \,+\, k\,e^{-nd_n(b+\delta)}, \tag{51}$$

where

$$\bar{J}_n(x) := \frac{1-e^{-n(d_n'(b)-x)\delta}}{n(d_n'(b)-x)} e^{-nd_n(b)}.$$

At this stage, denote

$$J_n(x) := \frac{e^{-nd_n(b)}}{n(d_n'(b)-x)}.$$

Fix $x < (\theta_0 - b)f'(b)/\{2(1-b)\}$, so that $x < d_n'(b)$ for sufficiently large n. If n is sufficiently large, then $(1-\varepsilon)J_n(x) < \bar{J}_n(x) < J_n(x)$. Hence, (51) becomes

$$(p(b)-\varepsilon)\,(1-\varepsilon)\,J_n(-\varepsilon) \le J_n \le (p(b)+\varepsilon)\,J_n(\varepsilon) \,+\, k\,e^{-nd_n(b+\delta)}. \tag{52}$$

Since d_n is increasing on $(b+\delta, 1) \subset (\theta_0, 1)$, $e^{-nd_n(b+\delta)} = o(J_n(x))$ as $n \to \infty$ for $x < (\theta_0 - b)f'(b)/\{2(1-b)\}$. In fact, $d_n(b)$, $d_n'(b)$ and $d_n(b+\delta)$ converge to positive constants by the strong law of large numbers, f (and therefore d_n and d_n') being continuous. Hence, (52) yields

$$(p(b)-\varepsilon)(1-\varepsilon)\,J_n(-\varepsilon) \le J_n \le (p(b)+2\varepsilon)\,J_n(\varepsilon)$$

for sufficiently large n. The number ε being arbitrary, it follows that

$$J_n \sim p(b)\,J_n(0)$$

as $n \to \infty$ and (46) is proved.

In order to prove (47), take $\bar{d_n}(t) := d(t,\,1-\hat{\theta}_n) = d_n(1-t)$, $\bar{p}(t) := p(1-t)$, $\bar{\theta}_0 := 1-\theta_0$ $\bar{b} := 1-a$ (so that $\bar{b} > \bar{\theta}_0$) and notice that by (46)

$$\int_{(\bar{b},\,1)} \bar{p}(s)\,e^{-n\,\bar{d_n}(s)}\,\mathrm{d}s \sim \frac{1}{n}\,\frac{\bar{p}(\bar{b})}{\bar{d_n}'(\bar{b})}\,e^{-n\,\bar{d_n}(\bar{b})},$$

and then apply the substitution $t = 1-s$ in the integral.

□

Proof of Proposition 2. Let $p = \mathrm{d}\pi/\mathrm{d}\lambda$, where λ is the Lebesgue measure. In order to apply Lemma 7, notice that

$$\pi_f^{(n)}((\theta_0 - \varepsilon, \theta_0 + \varepsilon)^c) = \frac{\int_{(0,\theta_0-\varepsilon)} e^{-n\, d(\theta,\hat{\theta}_n)} p(\theta)\,\mathrm{d}\theta + \int_{(\theta_0+\varepsilon,1)} e^{-n\, d(\theta,\hat{\theta}_n)} p(\theta)\,\mathrm{d}\theta}{\int_{(0,1)} e^{-n\, d(\theta,\hat{\theta}_n)} p(\theta)\,\mathrm{d}\theta}.$$

Hence, combining (43) with (47) and (46), the thesis follows.

□

Proof of Proposition 3. Denote by $p(\cdot)$ the density of π with respect to the Lebesgue measure. To begin, notice that

$$\begin{aligned}\int_{(0,1)} (t-\hat{\theta}_n)^2\, \pi_f^{(n)}(\mathrm{d}t) &= \frac{\int_{(0,1)} e^{-n\, d(t,\hat{\theta}_n)}\,(t-\hat{\theta}_n)^2\,\mathrm{d}\pi(t)}{\int_{(0,1)} e^{-n\, d(t,\hat{\theta}_n)}\,\mathrm{d}\pi(t)} \\ &= \frac{\int_{(0,1)} p(t)\, e^{-n\, d(t,\hat{\theta}_n)}\,(t-\hat{\theta}_n)^2\,\mathrm{d}t}{\int_{(0,1)} p(t)\, e^{-n\, d(t,\hat{\theta}_n)}\,\mathrm{d}t}.\end{aligned} \tag{53}$$

Lemma 7 can be applied for both the numerator and the denominator of (53). Combination of (53) with (43) and (44) yields

$$\int_{(0,1)} (t-\hat{\theta}_n)^2\, \pi_f^{(n)}(\mathrm{d}t) \sim \frac{1}{n\, d_n''(\hat{\theta}_n)}, \tag{54}$$

as $n \to \infty$.

Combining (54) with (12), one obtains that

$$\int_{(0,1)} (t-\hat{\theta}_n)^2\, \pi_f^{(n)}(\mathrm{d}t) \sim -\frac{1-\hat{\theta}_n}{n\, f'(\hat{\theta}_n)}, \tag{55}$$

as $n \to \infty$.

In virtue of continuity of f', by the strong law of large numbers, (55) entails that

$$\int_{(0,1)} (t-\hat{\theta}_n)^2\, \pi_f^{(n)}(\mathrm{d}t) \sim -\frac{1-\theta_0}{n f'(\theta_0)}, \tag{56}$$

as $n \to \infty$.

Let $E_f(\hat{\theta}_n)$ denote the the mean with respect to the distribution $\pi_f^{(n)}$.

Notice that

$$V_f(\hat{\theta}_n) = \int_{(0,1)} (t-\hat{\theta}_n)^2\, \pi_f^{(n)}(\mathrm{d}t) \;-\; (E_f(\hat{\theta}_n) - \hat{\theta}_n)^2, \tag{57}$$

and

$$\int_{(0,1)} (t-\hat{\theta}_n)^2\, \pi_f^{(n)}(\mathrm{d}t) = \int_{(0,1)} t^2\, \pi_f^{(n)}(\mathrm{d}t) \;-\; 2\hat{\theta}_n \int_{(0,1)} t\, \pi_f^{(n)}(\mathrm{d}t) \;+\; \hat{\theta}_n^2. \tag{58}$$

By (58), we obtain that

$$\begin{aligned}\left|E_f(\hat\theta_n)-\hat\theta_n\right| &= \frac{1}{2\hat\theta_n}\left|\int_{(0,1)} t^2\,\pi_f^{(n)}(\mathrm{d}t)-\hat\theta_n^2-\int_{(0,1)}(t-\hat\theta_n)^2\,\pi_f^{(n)}(\mathrm{d}t)\right|\\ &\le \frac{1}{2\hat\theta_n}\left(\left|\int_{(0,1)} t^2\,\pi_f^{(n)}(\mathrm{d}t)-\hat\theta_n^2\right|+\int_{(0,1)}(t-\hat\theta_n)^2\,\pi_f^{(n)}(\mathrm{d}t)\right)\end{aligned} \tag{59}$$

At this stage, dividing (45) by (43) and applying (12) and the strong law of large numbers, one obtains that

$$\int_{(0,1)} t^2\,\pi_f^{(n)}(\mathrm{d}t)-\hat\theta_n^2 \sim -\frac{1-\theta_0}{nf'(\theta_0)}. \tag{60}$$

In virtue of (56) and (60), equation (59) yields:

$$E_f(\hat\theta_n)-\hat\theta_n = O\Big(\frac{1}{n}\Big).$$

Hence, $(E_f(\hat\theta_n)-\hat\theta_n)^2$ is negligible with respect to (56) and therefore (57) entails that

$$V_f(\hat\theta_n)\sim\int_{(0,1)}(t-\hat\theta_n)^2\,\pi_f^{(n)}(\mathrm{d}t) \tag{61}$$

The thesis follows from (61) and (56). □

Acknowledgements

We thank two anonymous referees for their helpful comments. This work was partially supported by ESF and Regione Lombardia (by the grant "Dote Ricercatori").

References

Ali, S. M. and Silvey, S. D. (1966), 'A general class of coefficients of divergence of one distribution from another', *J. Roy. Statist. Soc. Ser. B* **28**(1), 131–142.

Berger, J. O. (1993), *Statistical decision theory and Bayesian analysis*, Springer-Verlag.

Bernardo, J. and Smith, A. (1994), *Bayesian theory*, Wiley.

Bissiri, P. and Walker, S. (2009), Converting information into probability measures using the Kullback–Leibler divergence, Technical report. Submitted.

Csiszár, I. (1967), 'Information-type measures of difference of probability distributions and indirect observation', *Studia Sci. Math. Hungar.* **2**, 229–318.

de Bruijn, N. (1981), *Asymptotic Methods in Analysis*, Dover Publications.

Kullback, S. and Leibler, R. (1951), 'On information and sufficiency', *Ann. Math. Statist.* **22**, 79–86.

Walker, S. G. (2006), 'Bayesian inference via a minimization rule', *Sankhyā* **68**, 542–553.

Xing, Y. and Ranneby, B. (2009), 'Sufficient conditions for Bayesian consistency', *J. Statist. Plann. Inference* **139**(7), 2479–2489.